Chapter 7

STABILITY OF LINEAR SYSTEMS

A general definition of stability is neither simple nor universal and depends on the particular phenomenon being considered. In the case of the nonlinear ODE's of interest in fluid dynamics, stability is often discussed in terms of fixed points, attractors, and the possibility of chaos. In these terms a system is said to be stable in a certain domain if, from within that domain, some norm of its solution is always attracted to the same fixed point. These are important and interesting concepts but we do not dwell on them in this work. Our basic concern is with time-dependent equations that are controlled by *stationary* systems, i.e., ODE's or OΔE's that are both linear and essentially autonomous, see Section 4.3.1. Chapters 4 and 6 developed the representative forms of ODE's generated from the basic PDE's by the semidiscrete approach, and then the OΔE's generated from the representative ODE's by application of time-marching methods. Stationary forms of these equations are represented by

$$\frac{d\vec{u}}{dt} = A\vec{u} - \vec{f}(t) \tag{7.1}$$

and

$$\vec{u}_{n+1} = \mathrm{C}\vec{u}_n - \vec{g}_n \tag{7.2}$$

respectively.

7.1 Dependence on the Eigensystem

Our definitions of stability are based entirely on the behavior of the homogeneous parts of Eqs. 7.1 and 7.2. The stability of Eq. 7.1 depends entirely on the eigensys-

tem[1] of A. The stability of Eq. 7.2 can often also be related to the eigensystem of its matrix. However, in this case the situation is not quite so simple since, in our applications to partial differential equations (especially hyperbolic ones), a stability definition can depend on both the time and space differencing. This is discussed in Section 7.4. Analysis of these eigensystems has the important added advantage that it gives an estimate of the *rate* at which a solution approaches a steady-state if a system is stable. Consideration will be given to matrices that have both complete and defective eigensystems, see Section 4.3.3, with a reminder that a complete system can be arbitrarily close to a defective one, in which case practical applications can make the properties of the latter appear to dominate.

If A and C are stationary, we can, in theory at least, estimate their fundamental properties. For example, in Section 4.4.2 we found from our model ODE's for diffusion and periodic convection what could be expected for the eigenvalue spectrums of practical physical problems containing these phenomena, see Fig. 4.1. These expectations are referred to many times in the following analysis of stability properties. They are important enough to be summarized by the following:

- For diffusion dominated flows the λ-eigenvalues tend to lie along the negative real axis.
- For periodic convection-dominated flows the λ-eigenvalues tend to lie along the imaginary axis.

In many interesting cases, the eigenvalues of the matrices in Eqs. 7.1 and 7.2 are sufficient to determine the stability. In previous chapters we designated these eigenvalues as λ_m and σ_m for Eqs. 7.1 and 7.2, respectively, and we will find it convenient to examine the stability of various methods in both the complex λ and complex σ planes.

7.2 Inherent Stability of ODE's

7.2.1 The Criterion

Here we state the standard stability criterion used for ordinary differential equations.

For a *stationary* matrix A, Eq. 7.1 is *inherently stable* if, when $\vec{f}$ is constant, $\vec{u}$ remains bounded as $t \to \infty$.	(7.3)

Note that inherent stability depends only on the transient solution of the ODE's.

[1]This is *not* the case for a nonautonomous system even if it is linear.

7.2.2 Complete Eigensystems

If a matrix has a complete eigensystem, all of its eigenvectors are linearly independent, and the matrix can be diagonalized by a similarity transformation. In such a case it follows at once from Eq. 6.24, for example, that the ODE's are inherently stable if and only if

$$\boxed{\Re(\lambda_m) \leq 0 \quad \text{for all} \quad m} \tag{7.4}$$

This states that, for inherent stability, all of the λ eigenvalues must lie on, or to the left of, the imaginary axis in the complex λ plane. Inspecting Fig. 4.1, we see that this criterion is satisfied for the model ODE's representing both diffusion and biconvection. It should be emphasized (as it is an important practical consideration in convection-dominated systems) that the special case for which $\lambda = \pm i$ is *included* in the domain of stability. In this case it is true that $\vec{u}$ does not decay as $t \to \infty$, but neither does it grow, so the above condition is met. Finally we note that for ODE's with complete eigensystems the eigenvectors play no role in the inherent stability criterion.

7.2.3 Defective Eigensystems

In order to understand the stability of ODE's that have defective eigensystems, we inspect the nature of their solutions in eigenspace. For this we draw on the results in Sections 4.3.3 and especially on Eqs. 4.18 to 4.19 in that section. In an eigenspace related to defective systems the form of the representative equation changes from a single equation to a Jordan block. For example, instead of Eq. 4.33 a typical form of the homogeneous part might be

$$\begin{bmatrix} u_1' \\ u_2' \\ u_3' \end{bmatrix} = \begin{bmatrix} \lambda & & \\ 1 & \lambda & \\ & 1 & \lambda \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

for which one finds the solution

$$\begin{aligned} u_1(t) &= u_1(0)e^{\lambda t} \\ u_2(t) &= [u_2(0) + u_1(0)t]e^{\lambda t} \\ u_3(t) &= \left[u_3(0) + u_2(0)t + \frac{1}{2}u_1(0)t^2\right]e^{\lambda t} \end{aligned} \tag{7.5}$$

Inspecting this solution, we see that for such cases condition 7.4 must be modified to the form

$$\boxed{\Re(\lambda_m) < 0 \quad \text{for all} \quad m} \tag{7.6}$$

since for pure imaginary λ, u_2 and u_3 would grow without bound (linearly or quadratically) if u_2 and u_1 were initially non-zero. Theoretically this condition is sufficient for stability in the sense of Statement 7.3 since $t^k e^{-|\epsilon|t} \to 0$ as $t \to \infty$ for all non-zero ϵ. However, in practical applications the criterion may be worthless since there may be a very large growth of the polynomial before the exponential "takes over" and brings about the decay. Furthermore, on a computer such a growth might destroy the solution process before it could be terminated.

Note that the stability condition 7.6 *excludes* the imaginary axis which tends to be occupied by the eigenvalues related to biconvection problems. However, condition 7.6 is of little or no practical importance if significant amounts of dissipation are present.

7.3 Numerical Stability of OΔE 's

7.3.1 The Criterion

The OΔE companion to Statement 7.3 is

$$\boxed{\text{For a } \textit{stationary} \text{ matrix C, Eq. 7.2 is } \textit{numerically stable} \text{ if, when } \vec{g} \text{ is constant, } \vec{u} \text{ remains bounded as } n \to \infty.} \tag{7.7}$$

We see that numerical stability depends only on the transient solution of the OΔE 's. This definition of stability is sometimes referred to as asymptotic or time stability.

As we stated at the beginning of this chapter, stability definitions are not unique. A definition often used in CFD literature stems from the development of PDE solutions that do not necessarily follow the semidiscrete route. In such cases it is appropriate to consider simultaneously the effects of both the time and space approximations. A time-space domain is fixed and stability is defined in terms of what happens to some norm of the solution within this domain as the mesh intervals go to zero at some constant ratio. We discuss this point of view in Section 7.4.

7.3.2 Complete Eigensystems

Consider a set of OΔE 's governed by a complete eigensystem. The stability criterion, according to the condition set in Eq. 7.7, follows at once from a study of Eq. 6.25 and its companion for multiple σ-roots, Eq. 6.31. Clearly, for such systems a time-marching method is numerically stable if and only if

$$\boxed{|(\sigma_m)_k| \leq 1 \quad \text{for all } m \text{ and } k} \tag{7.8}$$

This condition states that, for numerical stability, all of the σ eigenvalues (both principal and spurious, if there are any) must lie on or inside the unit circle in the complex σ-plane.

The similarity with the ODE discussion is very great. Again the sensitive case occurs for the periodic-convection model which places the "correct" location of the principal σ–root precisely on the unit circle where the solution is only neutrally stable. Further, for a complete eigensystem, the eigenvectors play no role in the numerical stability assessment.

7.3.3 Defective Eigensystems

The discussion for these systems parallels the discussion for defective ODE's. Examine Eq. 6.15 and note its similarity with Eq. 7.5. We see that for defective OΔE's the required modification to 7.8 is

$$\boxed{|(\sigma_m)_k| < 1 \quad \text{for all } m \text{ and } k} \tag{7.9}$$

since defective systems do not guarantee boundedness for $|\sigma| = 1$, for example in Eq. 7.5 if $|\sigma| = 1$ and $u_2(0)$ and/or $u_1(1) = 0$ we get linear or quadratic growth.

7.4 Time-Space Stability and Convergence of OΔE's

Let us now examine the concept of stability in a different way. In the previous discussion we considered in some detail the following approach:

1. The PDE's are converted to ODE's by approximating the space derivatives on a finite mesh.
2. Inherent stability of the ODE's is established by guaranteeing that $\mathrm{Re}(\lambda) \leq 0$.
3. Time-march methods are developed which guarantee that $|\sigma(\lambda h)| \leq 1$ and this is taken to be the condition for numerical stability.

This *does* guarantee that a linear autonomous system, generated from a PDE on some *fixed* space mesh, will have a numerical solution that is bounded as $t = nh \to \infty$. This *does not* guarantee that desirable solutions are generated in the time march process as both the time and space mesh intervals approach zero.

Now let us define stability in the time-space sense. First construct a finite time-space domain lying within $0 \leq x \leq L$ and $0 \leq t \leq T$. Cover this domain with a grid that is equispaced in both time and space and fix the mesh *ratio* by the equation

$$\Delta t = \Delta x \cdot c_n$$

Next reduce our OΔE approximation of the PDE to a two-level[2] (i.e., two time-planes) formula. Represent the homogeneous part of this formula by

$$\vec{u}_{n+1} = \mathrm{C}\vec{u}_n \tag{7.10}$$

Eq. 7.10 is said to be stable if any bounded initial vector, $\vec{u}_0$, produces a bounded solution vector, $\vec{u}_n$, as the mesh shrinks to zero for a fixed c_n. This is the classical definition of stability. It is often referred to as Lax or Lax-Richtmyer stability. Clearly as the mesh intervals go to zero, the number of time steps, N, must go to infinity in order to cover the entire fixed domain, so the criterion in 7.7 is a necessary condition for this stability criterion.

The significance of this definition of stability arises through *Lax's Theorem*, which states that, if a numerical method is *stable* (in the sense of Lax) and *consistent* then it is *convergent*. Consistency was briefly mentioned in Chapter 6 and is further discussed in Section 7.8. A method is *consistent* if it produces no error (in the Taylor series sense) in the limit as the mesh spacing and the time step go to zero (with c_n fixed, in the hyperbolic case). A method is *convergent* if it converges to the exact solution as the mesh spacing and time step go to zero in this manner. Clearly, this is an important property.

Applying simple recursion to Eq. 7.10, we find

$$\vec{u}_n = \mathrm{C}^n\vec{u}_0$$

and using vector and matrix p-norms (see Appendix 13.6) and their inequality relations, we have

$$||\vec{u}_n|| = ||\mathrm{C}^n\vec{u}_0|| \le ||\mathrm{C}^n|| \cdot ||\vec{u}_0|| \le ||\mathrm{C}||^n \cdot ||\vec{u}_0|| \tag{7.11}$$

Since the initial data vector is bounded, the solution vector is bounded if

$$||\mathrm{C}|| \le 1 \tag{7.12}$$

where $||\mathrm{C}||$ represents any p-norm of C. This is often used as a *sufficient* condition for stability.

Now we need to relate the stability definitions given in Eqs. 7.8 and 7.9 with that given in Eq. 7.12. In Eqs. 7.8 and 7.9, stability is related to the *spectral radius* of C, i.e., its eigenvalue of maximum magnitude. In Eq. 7.12, stability is related to a p-norm of C. It is clear that *the criteria are the same when the spectral radius is a true p-norm.*

Two facts about the relation between spectral radii and matrix norms are well known:

[2] Higher level equations can always be transformed to a two-level set by introducing new dependent variables.

1. The spectral radius of a matrix is its L_2 norm when the matrix is normal, i.e., it commutes with its transpose.

2. The spectral radius is the *lower bound* of all norms.

Furthermore, when C is normal, the second inequality in Eq. 7.11 becomes an equality. In this case, Eq. 7.12 becomes both necessary and sufficient for stability. From these relations we draw two important conclusions about the numerical stability of methods used to solve PDE's.

- The stability criteria in Eqs. 7.8 and 7.12 are identical for stationary systems when the governing matrix is normal. This includes symmetric, asymmetric, and circulant matrices. These criteria are both necessary and sufficient for methods that generate such matrices and depend solely upon the eigenvalues of the matrices.

- If the spectral radius of *any* governing matrix is greater than one, the method is unstable by any criterion. Thus for general matrices, the spectral radius condition is necessary but not sufficient for Lax-stability.

7.5 Numerical Stability Concepts in the Complex σ-Plane

7.5.1 σ-Root Traces Relative to the Unit Circle

Whether or not the semi-discrete approach was taken to find the differencing approximation of a set of PDE's, the final difference equations can be represented by

$$\vec{u}_{n+1} = \mathrm{C}\vec{u}_n - \vec{g}_n$$

Furthermore if C has a complete[3] eigensystem, the solution to the homogeneous part can always be expressed as

$$\vec{u}_n = c_1\sigma_1^n\vec{x}_1 + \cdots + c_m\sigma_m^n\vec{x}_m + \cdots + c_M\sigma_M^n\vec{x}_M$$

where the σ_m are the eigenvalues of C. If the semi-discrete approach *is* used, we can find a relation between the σ and the λ eigenvalues. Strictly speaking this relation is of no consequence in this Section. However, it serves as a very convenient guide as to where we might expect the σ-roots to lie relative to the unit circle in the complex

[3]The subject of defective eigensystems has been addressed. From now on we will omit further discussion of this special case.

σ-plane. For this reason we will proceed to trace the locus of the σ-roots as a function of the parameter λh for the equations modeling dissipation and periodic convection[4].

Locus of the exact trace

Figure 1 shows the exact trace of the σ-root if it is generated by either $e^{-|\lambda|h}$ or $e^{i\omega h}$. In both cases the $\bullet$ represents the starting value where $h = 0$ and $\sigma = 1$. As the magnitude of λh increases, the trace representing the dissipation model heads towards the origin where it resides when $\lambda h =\rightarrow -\infty$. On the other hand, as the magnitude of ωh increases, the trace representing the biconvection model spins around the circumference of the unit circle which it never leaves. We must be careful in interpreting σ when it is representing $e^{i\omega h}$. The fact that it *lies on* the unit circle means only that the *amplitude* of the representation is correct, it tells us nothing of the *phase* error (see Eq. 6.34). The phase error relates to the *position on* the unit circle.

Examples of some methods

Now let us compare the exact σ-root traces with some that are produced by actual time-marching methods. Table 7.1 shows the λ-σ relations for a variety of methods. Figures 7.2 and 7.3 illustrate the results produced by various methods when they are applied to the model ODE's for diffusion and periodic-convection, Eqs. 4.4 and 4.5. It is implied that the behavior shown is typical of what will happen if the methods are applied to diffusion- (or dissipation-) dominated or periodic convection-dominated problems as well as what does happen in the model cases. Most of the important possibilities are covered by the illustrations.

[4]Or, if you like, the parameter h for fixed values of λ equal to -1 and i for the dissipation and biconvection cases, respectively.

1	$\sigma - 1 - \lambda h = 0$	Explicit Euler
2	$\sigma^2 - 2\lambda h\sigma - 1 = 0$	Leapfrog
3	$\sigma^2 - (1 + \frac{3}{2}\lambda h)\sigma + \frac{1}{2}\lambda h = 0$	AB2
4	$\sigma^3 - (1 + \frac{23}{12}\lambda h)\sigma^2 + \frac{16}{12}\lambda h\sigma - \frac{5}{12}\lambda h = 0$	AB3
5	$\sigma(1 - \lambda h) - 1 = 0$	Implicit Euler
6	$\sigma(1 - \frac{1}{2}\lambda h) - (1 + \frac{1}{2}\lambda h) = 0$	Trapezoidal
7	$\sigma^2(1 - \frac{2}{3}\lambda h) - \frac{4}{3}\sigma + \frac{1}{3} = 0$	2nd O Backward
8	$\sigma^2(1 - \frac{5}{12}\lambda h) - (1 + \frac{8}{12}\lambda h)\sigma + \frac{1}{12}\lambda h = 0$	AM3
9	$\sigma^2 - (1 + \frac{13}{12}\lambda h + \frac{15}{24}\lambda^2 h^2)\sigma + \frac{1}{12}\lambda h(1 + \frac{5}{2}\lambda h) = 0$	ABM3
10	$\sigma^3 - (1 + 2\lambda h)\sigma^2 + \frac{3}{2}\lambda h\sigma - \frac{1}{2}\lambda h = 0$	Gazdag
11	$\sigma - 1 - \lambda h - \frac{1}{2}\lambda^2 h^2 = 0$	RK2
12	$\sigma - 1 - \lambda h - \frac{1}{2}\lambda^2 h^2 - \frac{1}{6}\lambda^3 h^3 - \frac{1}{24}\lambda^4 h^4 = 0$	RK4
13	$\sigma^2(1 - \frac{1}{3}\lambda h) - \frac{4}{3}\lambda h\sigma - (1 + \frac{1}{3}\lambda h) = 0$	Milne 4th

Table 7.1. Some $\lambda - \sigma$ Relations

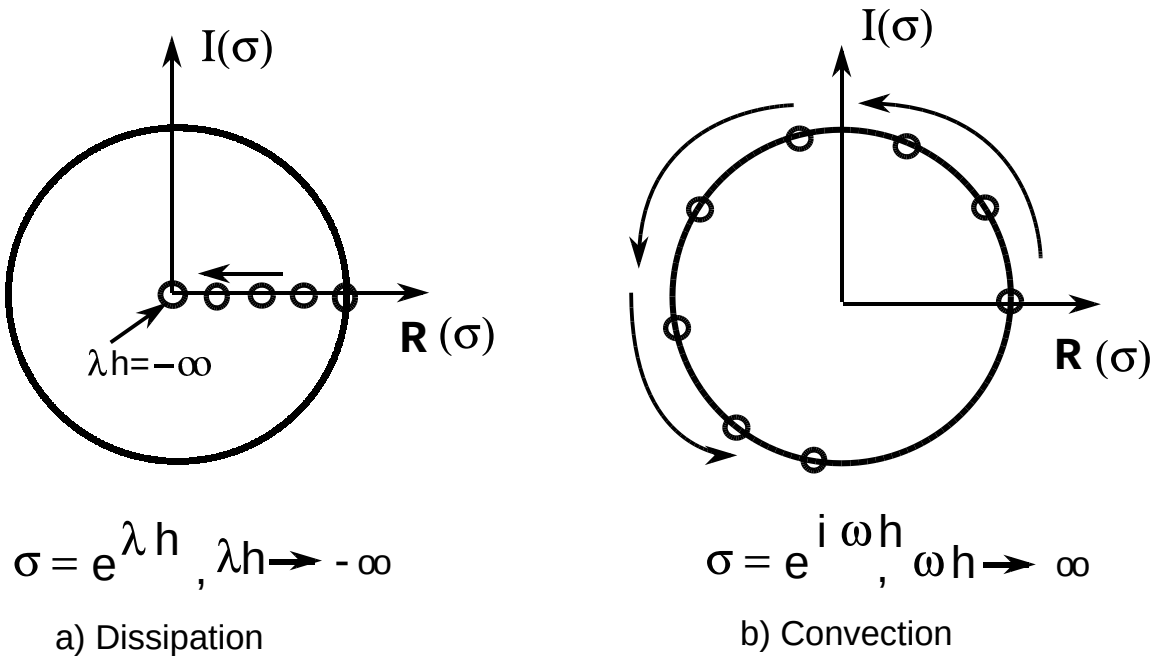

Figure 7.1: Ėxact traces of σ-roots for model equations.

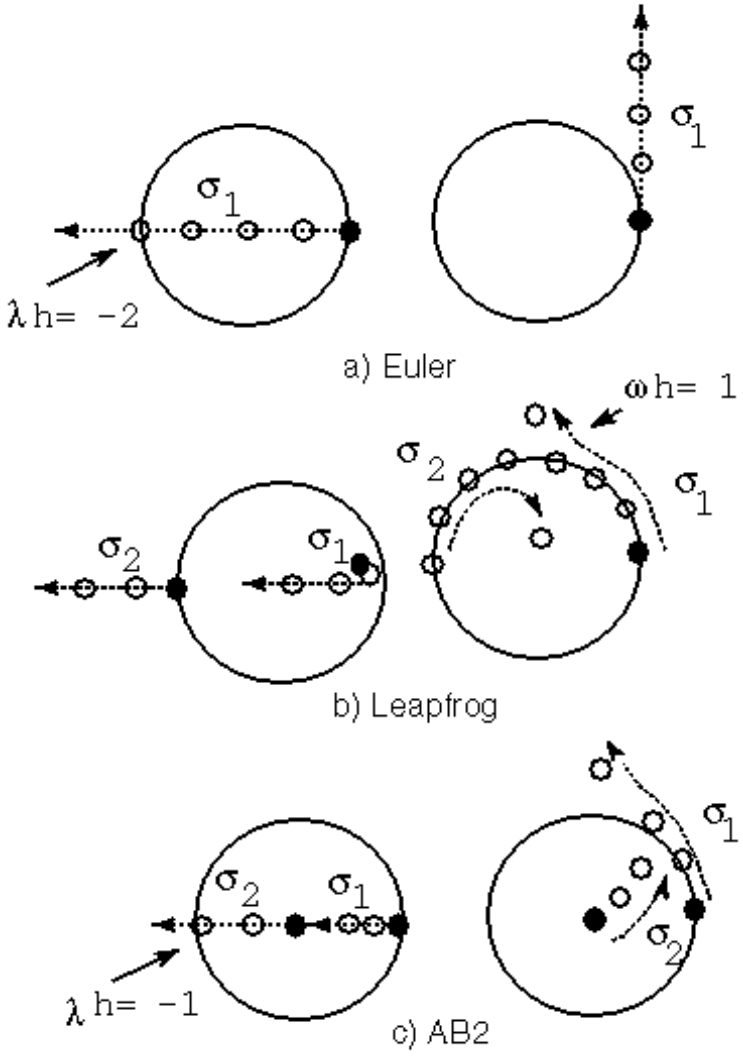

Figure 7.2: Traces of σ-roots for various methods.

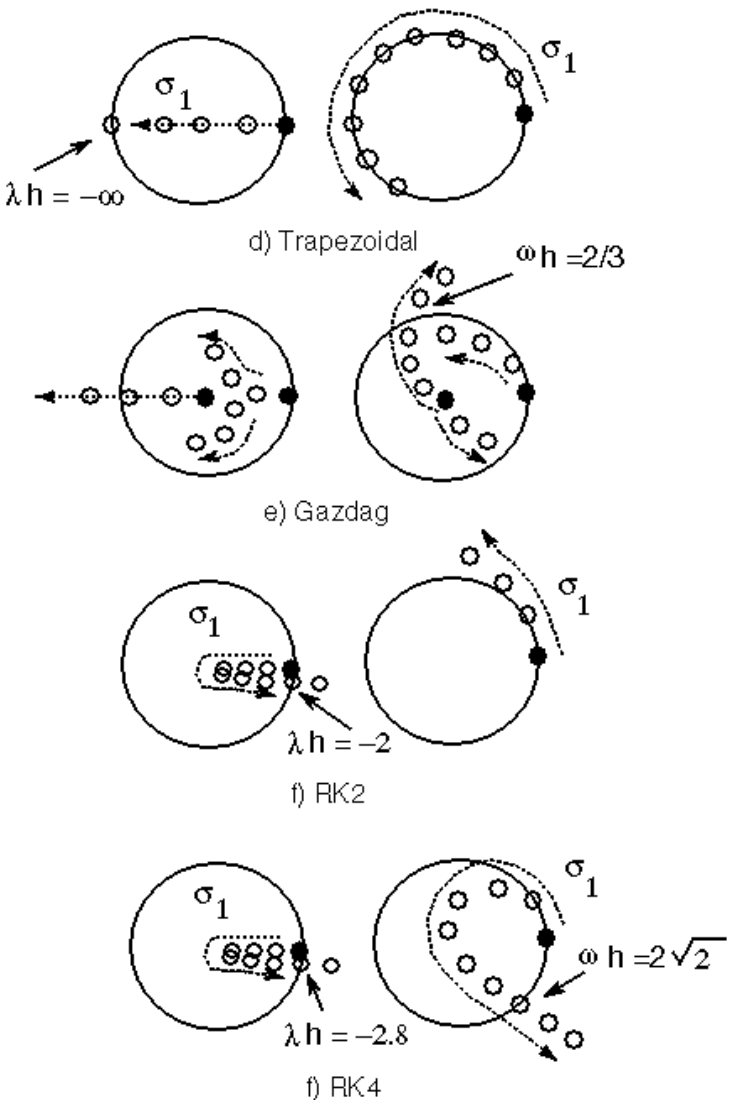

Figure 7.3: Traces of σ-roots for various methods (cont'd).

a. Explicit Euler Method

Figure 7.2 shows results for the explicit Euler method. When used for dissipation-dominated cases it is stable for the range -2$\leq \lambda h \leq$0. (Usually the magnitude of λ has to be estimated and often it is found by trial and error). When used for biconvection the σ- trace falls outside the unit circle for all finite h, and the method has no range of stability in this case.

b. Leapfrog Method

This is a two-root method, since there are two σ's produced by every λ. When applied to dissipation dominated problems we see from Fig. 7.2 that the principal root is stable for a range of λh, but the spurious root is not. In fact, the spurious root starts on the unit circle and falls outside of it for *all* $\Re(\lambda h) < 0$. However, for biconvection cases, when λ is *pure imaginary*, the method is not only stable, but it also produces a σ that falls precisely on the unit circle in the range $0 \leq \omega h \leq 1$. As was pointed out above, this does not mean, that the method is without error. Although the figure shows that there is a range of ωh in which the leapfrog method produces no error in amplitude, it says nothing about the error in *phase*. More is said about this in Chapter 8.

c. Second-Order Adams-Bashforth Method

This is also a two-root method but, unlike the leapfrog scheme, the spurious root starts at the origin, rather than on the unit circle, see Fig. 7.2. Therefore, there is a range of real negative λh for which the method will be stable. The figure shows that the range ends when $\lambda h < -1.0$ since at that point the spurious root leaves the circle and $|\sigma_2|$ becomes greater than one. The situation is quite different when the λ-root is pure imaginary. In that case as ωh increases away from zero the spurious root remains inside the circle and remains stable for a range of ωh. However, the principal root falls outside the unit circle for all $\omega h > 0$, and for the biconvection model equation the method is unstable for all h.

d. Trapezoidal Method

The trapezoidal method is a very popular one for reasons that are partially illustrated in Fig. 7.3. Its σ-roots fall on or inside the unit circle for both the dissipating and the periodic convecting case and, in fact, *it is stable for all values of* λh *for which* λ *itself is inherently stable*. Just like the leapfrog method it has the capability of producing only phase error for the periodic convecting case, but there is a major

difference between the two since the trapezoidal method produces no amplitude error for *any* ωh — not just a limited range between $0 \leq \omega h \leq 1$.

e. Gazdag Method

The Gazdag method was designed to produce low phase error. Since its characteristic polynomial for σ is a cubic (Table 7.1, no. 10), it must have two spurious roots in addition to the principal one. These are shown in Fig. 7.3. In both the dissipation and biconvection cases, a spurious root limits the stability. For the dissipating case, a spurious root leaves the unit circle when $\lambda h < -\frac{1}{2}$, and for the biconvecting case, when $\omega h > \frac{2}{3}$. Note that both spurious roots are located at the origin when $\lambda = 0$.

f,g. Second- and Fourth-Order Runge-Kutta Methods, RK2 and RK4

Traces of the σ-roots for the second- and fourth-order Runge-Kutta methods are shown in Fig. 7.3. The figures show that both methods are stable for a range of λh when λh is real and negative, but that the range of stability for RK4 is greater, going almost all the way to -2.8, whereas RK2 is limited to -2. On the other hand for biconvection RK2 is unstable for all ωh, whereas RK4 remains inside the unit circle for $0 \leq \omega h \leq 2\sqrt{2}$. One can show that the RK4 stability limit is about $|\lambda h| < 2.8$ for all complex λh for which $\Re(\lambda) \leq 0$.

7.5.2 Stability for Small Δt

It is interesting to pursue the question of stability when the time step size, h, is small so that accuracy of all the λ-roots is of importance. Situations for which this is not the case are considered in Chapter 8.

Mild instability

All conventional time-marching methods produce a principal root that is very close to $e^{\lambda h}$ for small values of λh. Therefore, on the basis of the principal root, the stability of a method that is required to resolve a transient solution over a relatively short time span may be a moot issue. Such cases are typified by the AB2 and RK2 methods when they are applied to a biconvection problem. Figs. 7.2c and 7.3f show that for both methods the principal root falls outside the unit circle and is unstable for all ωh. However, if the transient solution of interest can be resolved in a limited number of time steps that are small in the sense of the figure, the error caused by this instability may be relatively unimportant. If the root had fallen inside the circle the method would have been declared stable but an error of the same *magnitude* would have been committed, just in the opposite direction. For this reason the AB2 and the

RK2 methods have both been used in serious quantitative studies involving periodic convection. This kind of instability is referred to as mild instability and is not a serious problem under the circumstances discussed.

Catastrophic instability

There is a much more serious stability problem for small h that can be brought about by the existence of certain types of spurious roots. One of the best illustrations of this kind of problem stems from a critical study of the most accurate, explicit, two-step, linear multistep method (see Table 7.1):

$$u_{n+1} = -4u_n + 5u_{n-1} + 2h\left(2u'_n + u'_{n-1}\right) \tag{7.13}$$

One can show, using the methods given in Section 6.5, that this method is third-order accurate both in terms of er_λ and er_μ, so from an accuracy point of view it is attractive. However, let us inspect its stability even for very small values of λh. This can easily be accomplished by studying its characteristic polynomial when $\lambda h \to 0$. From Eq. 7.13 it follows that for $\lambda h = 0$, $P(E) = E^2 + 4E - 5$. Factoring $P(\sigma) = 0$ we find $P(\sigma) = (\sigma - 1)(\sigma + 5) = 0$. There are two σ-roots; σ_1, the principal one, equal to 1, and σ_2, a spurious one, equal to -5!!

In order to evaluate the consequences of this result, one must understand how methods with spurious roots work in practice. We know that they are not self starting, and the special procedures chosen to start them initialize the coefficients of the spurious roots, the c_{mk} for $k > 1$ in Eq. 6.31. If the starting process is well designed these coefficients are forced to be very small, and if the method is stable, they get smaller with increasing n. However, if the magnitude of one of the spurious σ is equal to 5, one can see disaster is imminent because $(5)^{10} \approx 10^7$. Even a very small initial value of c_{mk} is quickly overwhelmed. Such methods are called catastrophically unstable and are worthless for most, if not all, computations.

Milne and Adams type methods

If we inspect the σ-root traces of the multiple root methods in Figs. 7.2 and 7.3, we find them to be of two types. One type is typified by the leapfrog method. In this case a spurious root falls *on the unit circle* when $h \to 0$. The other type is exemplified by the 2nd-order Adams-Bashforth and Gazdag methods. In this case all spurious roots fall *on the origin* when $h \to 0$.

The former type is referred to as a *Milne Method*. Since at least one spurious root for a Milne method always starts on the unit circle, the method is likely to become unstable for some complex λ as h proceeds away from zero. On the basis of a Taylor

series expansion, however, these methods are generally the most accurate insofar as they minimize *the coefficient* in the leading term for er_t.

The latter type is referred to as an *Adams Method*. Since for these methods all spurious methods start at the origin for $h = 0$, they have a guaranteed range of stability for small enough h. However, on the basis of the magnitude of the coefficient in the leading Taylor series error term, they suffer, relatively speaking, from accuracy.

For a given amount of computational work, the order of accuracy of the two types is generally equivalent, and stability requirements in CFD applications generally override the (usually small) increase in accuracy provided by a coefficient with lower magnitude.

7.6 Numerical Stability Concepts in the Complex λh Plane

7.6.1 Stability for Large h.

The reason to study stability for small values of h is fairly easy to comprehend. Presumably we are seeking to resolve some transient and, since the accuracy of the transient solution for all of our methods depends on the smallness of the time-step, we seek to make the size of this step as small as possible. On the other hand, the cost of the computation generally depends on the number of steps taken to compute a solution, and to minimize this we wish to make the step size as large as possible. In the compromise, stability can play a part. Aside from ruling out catastrophically unstable methods, however, the situation in which all of the transient terms are resolved constitutes a rather minor role in stability considerations.

By far the most important aspect of numerical stability occurs under conditions when:

- One has inherently stable, coupled systems with λ–eigenvalues having widely separated magnitudes.

or

- We seek only to find a steady-state solution using a path that includes the unwanted transient.

In both of these cases there exist in the eigensystems relatively large values of $|\lambda h|$ associated with eigenvectors that we wish to drive through the solution process without any regard for their individual accuracy in eigenspace. This situation is the major motivation for the study of numerical stability. It leads to the subject of stiffness discussed in the next chapter.

7.6.2 Unconditional Stability, A-Stable Methods

Inherent stability of a set of ODE's was defined in Section 7.2 and, for coupled sets with a complete eigensystem, it amounted to the requirement that the real parts of all λ eigenvalues must lie on, or to the left of, the imaginary axis in the complex λ plane. This serves as an excellent reference frame to discuss and define the general stability features of time-marching methods. For example, we start with the definition:

> A numerical method is *unconditionally stable* if it is stable for all ODE's that are inherently stable.

A method with this property is said to be *A-stable*. A method is A_o *stable* if the region of stability contains the negative real axis in the complex λh plane, and *I-stable* if it contains the entire imaginary axis. By applying a fairly simple test for A-stability in terms of positive real functions to the class of two-step LMM's given in Section 6.48, one finds these methods to be A-stable if and only if

$$\theta \geq \varphi + \frac{1}{2} \tag{7.14}$$

$$\xi \geq -\frac{1}{2} \tag{7.15}$$

$$\xi \leq \theta + \varphi - \frac{1}{2} \tag{7.16}$$

A set of A-stable implicit methods is shown in Table 7.2.

θ	ξ	φ	Method	Order
1	0	0	Implicit Euler	1
1/2	0	0	Trapezoidal	2
1	1/2	0	2nd O Backward	2
3/4	0	−1/4	Adams type	2
1/3	−1/2	−1/3	Lees type	2
1/2	−1/2	−1/2	Two-step trapezoidal	2
5/8	−1/6	−2/9	A-contractive	2

Table 7.2. Some unconditionally stable (A-stable) implicit methods.

Notice that none of these methods has an accuracy higher than second-order. It can be proved that the order of an A-stable LMM *cannot exceed two*, and, furthermore that of all 2nd-order, A-stable methods the trapezoidal method has the smallest truncation error.

Returning to the stability test using positive real functions one can show that a two-step LMM is A_o-stable if and only if

$$\theta \geq \varphi + \frac{1}{2} \tag{7.17}$$

$$\xi \geq -\frac{1}{2} \tag{7.18}$$

$$0 \leq \theta - \varphi \tag{7.19}$$

For first-order accuracy, the inequalities 7.17 to 7.19 are less stringent than 7.14 to 7.16. For second-order accuracy, however, the parameters (θ, ξ, φ) are related by the condition

$$\varphi = \xi - \theta + \frac{1}{2}$$

and the two sets of inequalities reduce to the same set which is

$$\xi \leq 2\theta - 1 \tag{7.20}$$

$$\xi \geq -\frac{1}{2} \tag{7.21}$$

Hence, two-step, second-order accurate LMM's that are A-stable and A_o-stable share the same (φ, ξ, θ) parameter space. Although the order of accuracy of an A-stable method cannot exceed two, A_o-stable LMM methods exist which have an accuracy of arbitrarily high order.

It has been shown that for a method to be I-stable it must also be A-stable. Therefore, no further discussion is necessary for the special case of I-stability.

It is not difficult to prove that methods having a characteristic polynomial for which the coefficient of the highest order term in E is unity[5] can never be unconditionally stable. This includes all explicit methods and predictor-corrector methods made up of explicit sequences. Such methods are referred to, therefore, as *conditionally stable* methods.

7.6.3 Stability Contours in the Complex λh Plane.

A very convenient way to present the stability properties of a time-marching method is to plot the locus of the complex λh for which $|\sigma| = 1$, such that the resulting contour goes through the point $\lambda h = 0$. Here $|\sigma|$ refers to the maximum absolute value of any σ, principal or spurious, that is a root to the characteristic polynomial for

[5]Or can be made equal to unity by a trivial normalization (division by a constant independent of λh). The proof follows from the fact that the coefficients of such a polynomial are sums of various combinations of products of all its roots.

a given λh. It follows from Section 7.3 that on one side of this contour the numerical method is stable while on the other, it is unstable. We refer to it, therefore, as a *stability contour*.

Typical stability contours for both explicit and implicit methods are illustrated in Fig. 7.4, which is derived from the one-root θ-method given in Section 6.4.4.

Contours for explicit methods

Fig. 7.4a shows the stability contour for the explicit Euler method. In the following two ways it is typical of all stability contours for explicit methods:

1. The contour encloses a finite portion of the left-half complex λh-plane.
2. The region of stability is *inside* the boundary, and therefore, it is conditional.

However, this method includes no part of the imaginary axis (except for the origin) and so it is unstable for the model biconvection problem. Although several explicit methods share this deficiency (e.g., AB2, RK2), several others do not (e.g., leapfrog, Gazdag, RK3, RK4), see Figs. 7.5 and 7.6. Notice in particular that the third- and fourth-order Runge-Kutta methods, Fig. 7.6, include a portion of the imaginary axis out to $\pm 1.9i$ and $\pm 2\sqrt{2}i$, respectively.

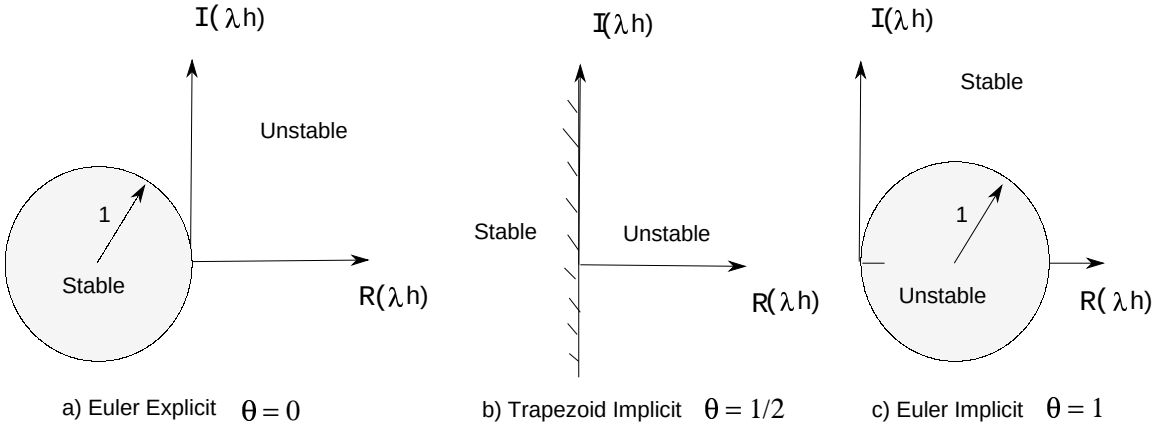

Figure 7.4: Ṡtability Contours for the θ-method.

Contours for unconditionally stable implicit methods

Fig. 7.4c shows the stability contour for the implicit Euler method. It is typical of many stability contours for unconditionally stable implicit methods. Notice that the

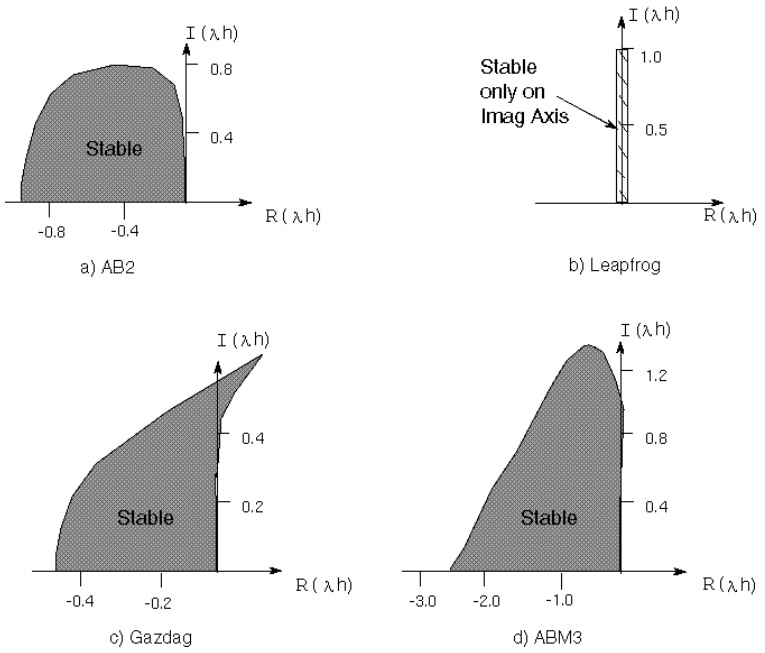

Figure 7.5: Ṡtability Contours for some explicit methods.

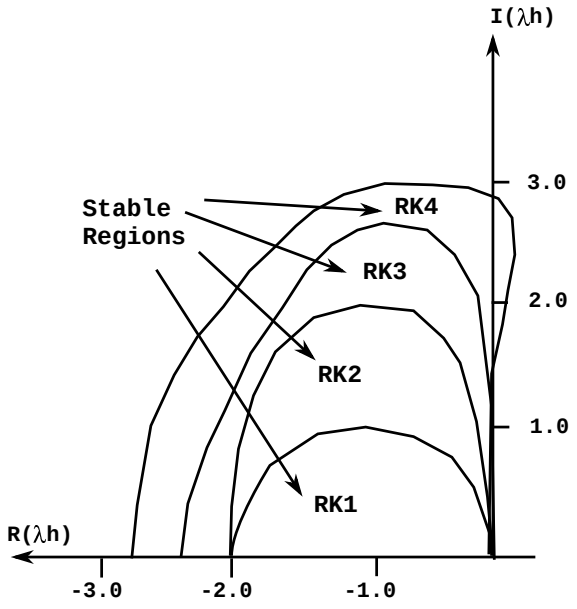

Figure 7.6: Ṡtability Contours for Runge-Kutta methods.

method is stable for the entire range of complex λh that fall *outside* the boundary. This means that the method is numerically stable *even when the ODE's that it is being used to integrate are inherently unstable*. Some other implicit unconditionally stable methods with the same property are shown in Fig. 7.7. In all of these cases the imaginary axis is part of the *stable* region.

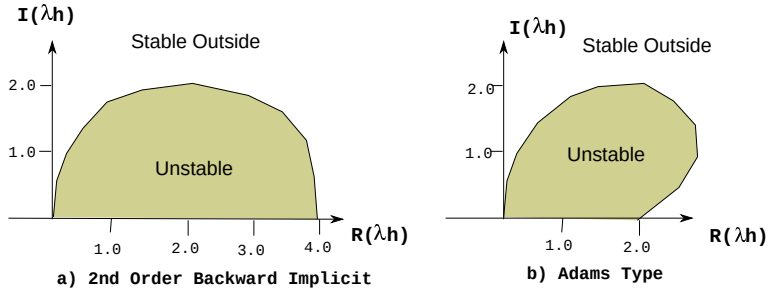

Figure 7.7: Stability Contours for the 2 unconditionally stable implicit methods.

Not all unconditionally stable methods are stable in some regions where the ODE's they are integrating are inherently unstable. The classic example of a method that is stable *only* when the generating ODE's are themselves inherently stable is the trapezoidal method, i.e., the special case of the θ-method for which $\theta = \frac{1}{2}$. The stability boundary for this case is shown in Fig. 7.4b. The boundary is the imaginary axis and the numerical method is stable for λh lying on or to the left of this axis. Two other methods that have this property are the two-step trapezoidal method

$$u_{n+1} = u_{n-1} + h\left(u'_{n+1} + u'_{n-1}\right)$$

and a method due to Lee

$$u_{n+1} = u_{n-1} + \frac{2}{3}h\left(u'_{n+1} + u'_n + u'_{n-1}\right)$$

Notice that both of these methods are of the Milne type.

Contours for conditionally stable implicit methods

Just because a method is implicit does not mean that it is unconditionally stable. Two illustrations of this are shown in Fig. 7.8. One of these is the Adams-Moulton 3rd-order method (no. 8, Table 7.1). Another is the 4th-order Milne method given by the point operator

$$u_{n+1} = u_{n-1} + \frac{1}{3}h\left(u'_{n+1} + 4u'_n + u'_{n-1}\right)$$

and shown in Table 7.1 as no. 13. It is stable *only* for $\lambda = \pm i\omega$ when $0 \leq \omega \leq \sqrt{3}$. Its stability boundary is very similar to that for the leapfrog method (see Fig. 7.5b).

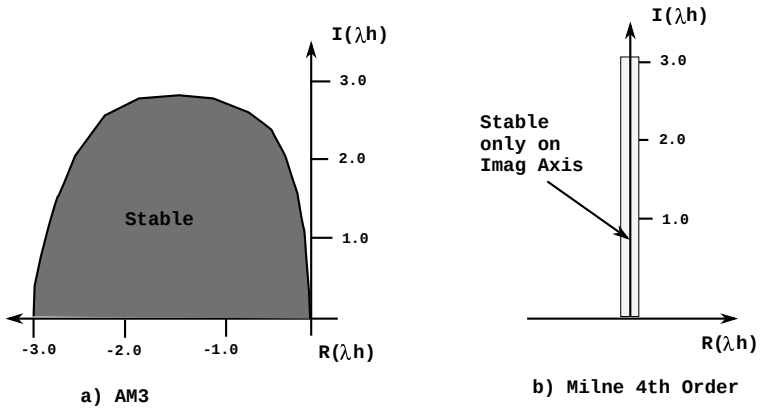

Figure 7.8: Ṡtability Contours for the 2 conditionally stable implicit methods.

7.7 Fourier or von Neumann Stability Analysis

By far the most popular form of stability analysis for numerical schemes is the Fourier or von Neumann approach. This analysis is usually carried out on point operators and it does not depend on an intermediate stage of ODE's. *Strictly speaking* it applies only to difference approximations of PDE's that produce OΔE's *which are linear, have no space or time varying coefficients, and have periodic boundary conditions.*[6] In practical application it is often used as a guide for estimating the worthiness of a method. Years of experience have shown that it serves as a fairly reliable *necessary* stability condition, but it is by no means a *sufficient* one.

7.7.1 The Basic Procedure

One takes data from a "typical" point in the flow field and uses this as constant throughout time and space according to the assumptions given above. Then one imposes a spatial harmonic as an initial value on the mesh and asks the question: Will its amplitude grow or decay in time? The answer is determined by finding the conditions under which

$$u(t,x) = e^{\alpha t} \cdot e^{ikx} \tag{7.22}$$

is a solution to the *difference* equation. Since, for the general term,

$$u_{j+m}^{(n+\ell)} = e^{\alpha(t+\ell\Delta t)} \cdot e^{ik(x+m\Delta x)} = e^{\alpha\ell\Delta t} \cdot e^{ikm\Delta x} \cdot u_j^{(n)}$$

[6]Another way of viewing this is to consider it as an initial value problem on an infinite space domain.

the quantity $u_j^{(n)}$ is common to every term and can be factored out. In the remaining expressions, we find the term $e^{\alpha\Delta t}$ which we represent by σ, thus:

$$\sigma \equiv e^{\alpha\Delta t}$$

Then, since $e^{\alpha t} = \left(e^{\alpha\Delta t}\right)^n = \sigma^n$, it is clear that

$$\boxed{\text{For numerical stability } |\sigma| \leq 1} \tag{7.23}$$

and the problem is to solve for the σ's produced by any given method and, as a necessary condition for stability, make sure that, in the worst possible combination of parameters, condition 7.23 is satisfied[7].

7.7.2 Some Examples

The procedure can best be explained by examples. Consider as a first example the finite difference approximation to the model diffusion equation known as Richardson's method of overlapping steps. This was mentioned in Section 4.1 and given as Eq. 4.1:

$$u_j^{(n+1)} = u_j^{(n-1)} + \nu\frac{2\Delta t}{\Delta x^2}\left(u_{j+1}^{(n)} - 2u_j^{(n)} + u_{j-1}^{(n)}\right) \tag{7.24}$$

Substitution of Eq. 7.22 into Eq. 7.24 gives the relation

$$\sigma = \sigma^{-1} + \nu\frac{2\Delta t}{\Delta x^2}\left(e^{ik\Delta x} - 2 + e^{-ik\Delta x}\right)$$

or

$$\sigma^2 + \underbrace{\left[\frac{4\nu\Delta t}{\Delta x^2}(1 - \cos k\Delta x)\right]}_{2b}\sigma - 1 = 0 \tag{7.25}$$

Thus Eq. 7.22 is a solution of Eq. 7.24 if σ is a root of Eq. 7.25. The two roots of 7.25 are

$$\sigma_{1,2} = -b \pm \sqrt{b^2 + 1}$$

from which it is clear that one $|\sigma|$ is always > 1. We find, therefore, that by the Fourier stability test, Richardson's method of overlapping steps is unstable for all ν, k and Δt.

[7] If boundedness is required in a *finite* time domain, the condition is often presented as $|\sigma| \leq 1 + O(\Delta t)$.

As another example consider the finite-difference approximation for the model biconvection equation

$$u_j^{(n+1)} = u_j^{(n)} - \frac{a\Delta t}{2\Delta x}\left(u_{j+1}^{(n)} - u_{j-1}^{(n)}\right) \tag{7.26}$$

In this case

$$\sigma = 1 - \frac{a\Delta t}{\Delta x} \cdot i \cdot \sin k\Delta x$$

from which it is clear that $|\sigma| > 1$ for all nonzero a and k. Thus we have another finite-difference approximation that, by the Fourier stability test, is unstable for any choice of the free parameters.

7.7.3 Relation to Circulant Matrices

The underlying assumption in a Fourier stability analysis is that the C matrix, determined when the differencing scheme is put in the form of Eq. 7.2, is circulant. Such being the case, the e^{ikx} in Eq. 7.22 represents an *eigenvector* of the system, and the two examples just presented outline a simple procedure for finding the *eigenvalues* of the circulant matrices formed by application of the two methods to the model problems. The choice of σ for the stability parameter in the Fourier analysis, therefore, is not an accident. It is exactly the same σ we have been using in all of our previous discussions, but arrived at from a different perspective.

If we examine the preceding examples from the viewpoint of circulant matrices and the semi-discrete approach, the results present rather obvious conclusions. The space differencing in Richardson's method produces the matrix $s \cdot B_p(1, -2, 1)$ where s is a positive scalar coefficient. From Appendix A.5 we find that the eigenvalues of this matrix are real negative numbers. Clearly, the time-marching is being carried out by the leapfrog method and, from Fig. 7.5, this method is unstable for *all* eigenvalues with negative real parts. On the other hand, the space matrix in Eq. 7.26 is $B_p(-1, 0, 1)$, and according to Appendix A.5, this matrix has pure imaginary eigenvalues. However, in this case the explicit Euler method is being used for the time-march and, according to Fig. 7.4, this method is always unstable for such conditions.

7.8 Consistency

Consider the model equation for diffusion analysis

$$\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} \tag{7.27}$$

Many years before computers became available (1910, in fact), Lewis F. Richardson proposed a method for integrating equations of this type. We presented his method in Eq. 4.1 and analyzed its stability by the Fourier method in Section 7.7.

In Richardson's time, the concept of numerical instability was not known. However, the concept is quite clear today and we now know immediately that his approach would be unstable. As a semi-discrete method it can be expressed in matrix notation as the system of ODE's:

$$\frac{d\vec{u}}{dt} = \frac{\nu}{\Delta x^2} B(1,-2,1)\vec{u} + (bc) \tag{7.28}$$

with the leapfrog method used for the time march. Our analysis in this Chapter revealed that this is numerically unstable since the λ-roots of B(1,-2,1) are all real and negative and the spurious σ-root in the leapfrog method is unstable for all such cases, see Fig. 7.5b.

The method was used by Richardson for weather prediction, and this fact can now be a source of some levity. In all probability, however, the hand calculations were not carried far enough to exhibit strange phenomena. We could, of course, use the 2nd-order Runge-Kutta method to integrate Eq. 7.28 since it is stable for real negative λ's. It is, however, conditionally stable and for this case we are rather severely limited in time step size by the requirement $\Delta t \leq \Delta x^2/(2\nu)$.

There are many ways to manipulate the numerical stability of algorithms. One of them is to introduce mixed time and space differencing, a possibility we have not yet considered. Let us investigate the following example

$$u_j^{(n+1)} = u_j^{(n-1)} + \frac{2\nu\Delta t}{\Delta x^2}\left[u_{j-1}^{(n)} - 2\left(\frac{u_j^{(n+1)} + u_j^{(n-1)}}{2}\right) + u_{j+1}^{(n)}\right]$$

in which the central term in the space derivative in Eq. 4.1 has been replaced by its average value at two different time levels. This was introduced as the DuFort-Frankel method in Chapter 4. Now let

$$\alpha \equiv \frac{2\nu\Delta t}{\Delta x^2} \tag{7.29}$$

and rearrange terms

$$(1+\alpha)u_j^{(n+1)} = (1-\alpha)u_j^{(n-1)} + \alpha\left[u_{j-1}^{(n)} + u_{j+1}^{(n)}\right]$$

There is no obvious ODE between the basic PDE and this final OΔE. Hence, there is no intermediate λ-root structure to inspect. Instead one proceeds immediately to the σ-roots.

The simplest way to carry this out is by means of the Fourier stability analysis introduced in Section 7.7. This leads at once to

$$(1+\alpha)\sigma = (1-\alpha)\sigma^{-1} + \alpha\left(e^{ik\Delta x} + e^{-ik\Delta x}\right)$$

or

$$(1+\alpha)\sigma_k^2 - 2\alpha\sigma_k \cos(k\Delta x) - (1-\alpha) = 0$$

where

$$\Delta x = \frac{2\pi}{M} \qquad \text{and} \qquad k = 1, 2, \cdots, M-1$$

The solution of the quadratic is

$$\sigma_k = \frac{\alpha\cos\theta_k \pm \sqrt{1-\alpha^2\sin^2\theta_k}}{1+\alpha}$$

where

$$\theta_k = \frac{2\pi k}{M} \qquad = 1, 2, \cdots, M-1$$

There are 2M σ-roots all of which are ≤ 1 for any real α in the range $0 \leq \alpha \leq \infty$. This means that the method is *unconditionally stable*!

The above result seems too good to be true, since we have found an unconditionally stable method using an *explicit* combination of Lagrangian interpolation polynomials.

> The price we have paid for this is the loss of *consistency* with the original PDE.

To prove this, we expand the terms in Eq. 7.29 in a Taylor series and reconstruct the partial differential equation we are actually solving as the mesh size becomes very small. For the time derivative we have

$$\frac{1}{2\Delta t}\left[u_j^{(n+1)} - u_j^{(n-1)}\right] = (\partial_t u)_j^{(n)} + \frac{1}{6}\Delta t^2 (\partial_{ttt} u)_j^{(n)} + \cdots$$

and for the mixed time and space differences

$$\begin{aligned}\frac{u_{j-1}^{(n)} - u_j^{(n+1)} - u_j^{(n-1)} + u_{j+1}^{(n)}}{\Delta x^2} &= (\partial_{xx} u)_j^{(n)} - \left(\frac{\Delta t}{\Delta x}\right)^2 (\partial_{tt} u)_j^{(n)} + \\ &\quad \frac{1}{12}\Delta x^2 (\partial_{xxxx} u)_j^{(n)} - \\ &\quad \frac{1}{12}\Delta t^2 \left(\frac{\Delta t}{\Delta x}\right)^2 (\partial_{tttt} u)_j^{(n)} + \cdots \qquad (7.30)\end{aligned}$$

Replace the terms in Eq. 7.29 with the above expansions and take the limit as Δt , $\Delta x \to 0$. We find

$$\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2} - \nu r^2 \frac{\partial^2 u}{\partial t^2} \tag{7.31}$$

where

$$r \equiv \frac{\Delta t}{\Delta x}$$

Eq. 7.27 is *parabolic*. Eq. 7.31 is *hyperbolic*. Thus if $\Delta t \to 0$ and $\Delta x \to 0$ in such a way that $\frac{\Delta t}{\Delta x}$ remains constant, the equation we actually solve by the method in Eq. 7.29 is a wave equation, not a diffusion equation. In such a case *Eq. 7.29 is not uniformly consistent with the equation we set out to solve* even for vanishingly small step sizes. The situation is summarized in Table 7.3.

2nd O Runge-Rutta		Du Fort–Frankel
	For Stability	
$\Delta t \leq \frac{\Delta x^2}{2\nu}$		$\Delta t \leq \infty$
Conditionally Stable		Unconditionally Stable
	For Consistency	
Uniformly Consistent		Conditionally Consistent,
With $\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2}$		Approximates $\frac{\partial u}{\partial t} = \nu \frac{\partial^2 u}{\partial x^2}$ only if $\nu \left(\frac{\Delta t}{\Delta x}\right)^2 < \epsilon$ Therefore $\Delta t < \Delta x \sqrt{\frac{\epsilon}{\nu}}$

Table 7.3: Summary of accuracy and consistency conditions for RK2 and Du Fort-Frankel methods. ϵ = an arbitrary error bound.